30 October 2020

OCLC Shared Entity Management Infrastructure

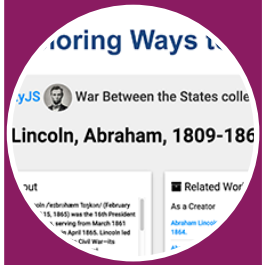
John Chapman

Senior Product Manager

Metadata Strategy and Operations

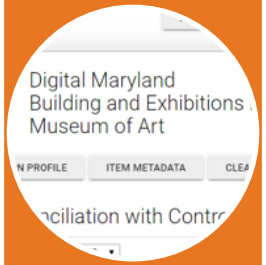
Building on our experience

Publish linked data - FAST, VIAF, WorldCat (2009 -)



EntityJS Research Project (2013)

Person Entity Lookup Pilot (2014)



CONTENTdm Metadata Refinery (2015-16)

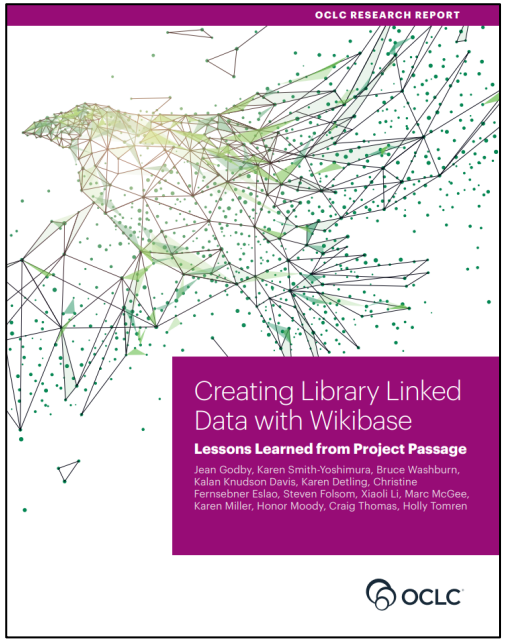
Project Passage (2017-18)

CONTENTdm Linked Data Pilot (2019-20)

Entity Management Infrastructure (2020-21)

OCLC

Feedback from Passage participants



oc.lc/passagereport

- Provide persistent identifiers relevant to library workflows
- Enable the creation of new identifiers within metadata management workflows
- Provide interfaces and ecosystem to create native linked data descriptions
- Seed the web with persistent identifiers
- Provide broad reconciliation across vocabularies & ontologies

OCLC awarded Mellon Foundation grant to develop infrastructure to support linked data management initiatives

'Entity Management Infrastructure' will advance use of linked data and ultimately improve discoverability of scholarly materials on the web

DUBLIN, Ohio, 9 January 2020—OCLC has been awarded a grant from The Andrew W. Mellon Foundation to develop a shared "Entity Management Infrastructure" that will support linked data management initiatives underway in the library and scholarly communications community. When complete, this infrastructure will be jointly curated by the community and OCLC, and will ultimately make scholarly materials more connected and discoverable on the web.

The two-year grant, for $2.436 million, will support work on the project that will run from January 2020 to December 2021. The Mellon grant funding represents approximately half of the total cost of the Entity Management Infrastructure project. OCLC is contributing the remaining half of the required investment.

"OCLC has been a leader in library linked data research for years, and we have developed prototypes, innovative pilot programs and partnerships that continue to inform our work," said Skip Prichard, OCLC President and CEO. "OCLC enables libraries to work together to achieve economies, efficiencies, and consistency in metadata creation. We're grateful to The Mellon Foundation for their generous support for this project. And we're eager to apply our knowledge and expertise to develop this infrastructure on behalf of libraries and the scholarly communications community."

THE ANDREW W. MELLON FOUNDATION

Visit oc.lc/mellon-grant to learn more

OCLC

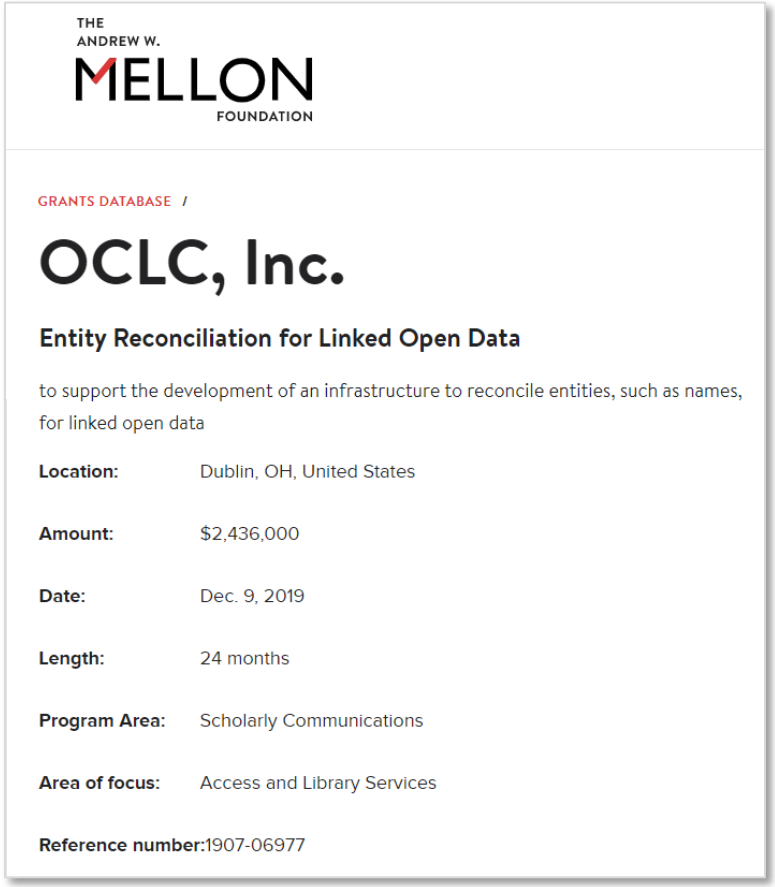
For linked data to move into common use, libraries need reliable and persistent identifiers and metadata for the critical entities they rely on. This project begins to build that infrastructure and advances the whole field.

Lorcan Dempsey
OCLC Vice President, Membership and Research, and Chief Strategist

Photo by Robynne Hu on Unsplash

Project overview

- Two-year, $2.436M grant, matched by OCLC
- Production infrastructure for Work and Person entities
- Support for multiple descriptive and encoding standards
- Use of persistent identifiers
- **Most importantly: a collaboration with the library community**



THE ANDREW W. MELLON FOUNDATION

GRANTS DATABASE /

OCLC, Inc.

Entity Reconciliation for Linked Open Data

to support the development of an infrastructure to reconcile entities, such as names, for linked open data

Location: Dublin, OH, United States

Amount: $2,436,000

Date: Dec. 9, 2019

Length: 24 months

Program Area: Scholarly Communications

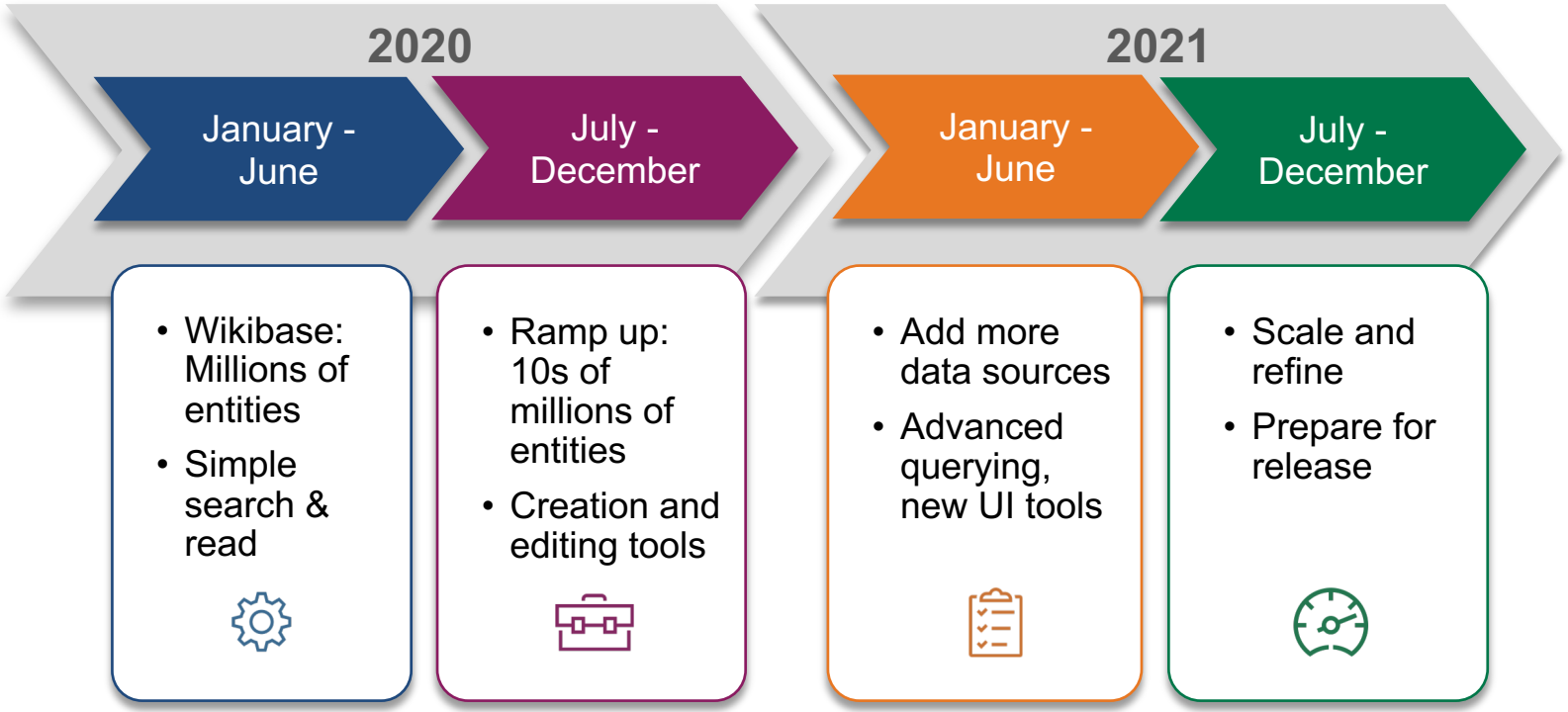
Area of focus: Access and Library Services

Reference number: 1907-06977

Our goals

- Address infrastructure needs identified by libraries
  - Expand on “native” metadata management
  - Link library data to non-library data… and shared data to local data
  - Provide ID creation services to help “at the point of need”
  - Stand behind entity URIs
- Operate at a large scale – and be sustainable
- Complement other efforts
- Deliver products and services December 2021

Timeline of activities



Production release

OCLC

What we have learned so far

- Need to increase capabilities for monitoring quality, breadth, depth
- APIs, machines as “users”
- Playing role in larger ecosystem
- Need redundancy, multiple environments, and robust testing capabilities
- Need to engineer loading and ingest technologies

Conversations continue

- Advisory Group
- LD4P, PCC
- Coffee chats
- Opening up testing outside of formal releases

Data activities

- Entity modeling
  - Works, persons, places
- Working at scale
- Quality composites

Minimum Viable Entity (MVE): Work

- MVE for Work
  - Some elements from Wikibase: label, description, also known as (when applicable)
  - Remaining elements based on LRM & BIBFRAME,
    i.e. a combination of Work and Expression elements: instance of, title, agent, realization date (often based on publication date for first known realization), content type, exemplar identifier (points to a thing in WorldCat)

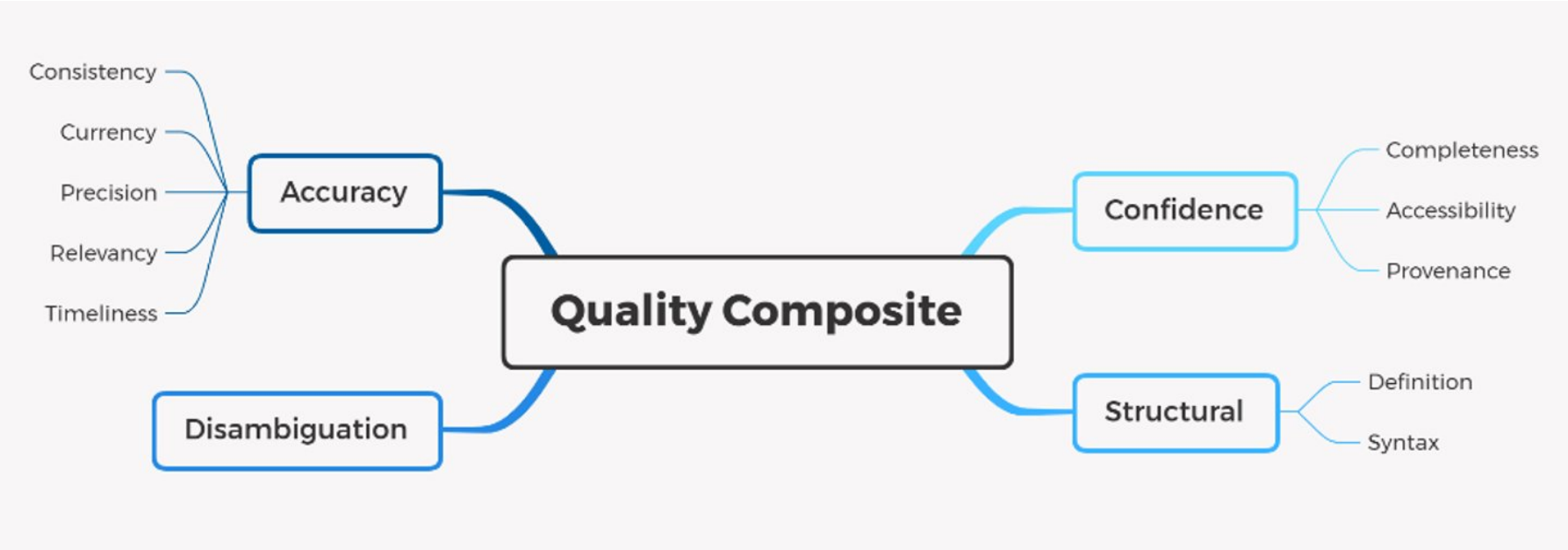
Processes

- Staff focused on two areas: modeling and quality
- For modeling work, documenting:
  - MVE description
  - SPARQL queries to validate MVE model
  - Data selection – sources, and logic used to select data

Scale

- Refining processes for ontology definition and data loading
- Goal is 100M+ entities by Dec 31
  - Roughly 80% works, 20% persons (<.01% places)

Quality Composite



What's next

- Continue to build out data models and entity descriptions
- Continue discussion of subscription models and pricing
- API rollout
- Development focus on creation and editing tools
- Broader testing

Questions?

OCLC